(12) UK Patent Application (19) GB (11) 2 213 623(13)A

(43) Date of A publication 16.08.1989

(21) Application No 8828532.5

(22) Date of filing 07.12.1988

(30) Priority data
(31) 62310569 (32) 08.12.1987 (33) JP
62323307 21.12.1987
62331656 25.12.1987

(51) INT CL⁴
G10L 5/06

(52) UK CL (Edition J)
G4R RPS R1F R11A R11D R11E R12F R3B R3G
R8F R8G R9B

(56) Documents cited
None

(58) Field of search
UK CL (Edition J) G4R RPD RPS RPW, H4R RPV
RPVA
INT CL⁴ G10L

(71) Applicant
Sony Corporation

(Incorporated in Japan)

6-7-35 Kitashinagawa, Shinagawa-ku, Tokyo 141,
Japan

(72) Inventors
Makoto Akabane
Yoichiro Sako
Atsunobu Hiraiwa

(74) Agent and/or Address for Service
D Young & Co
10 Staple Inn, London, WC1V 7RD, United Kingdom

(54) Phoneme recognition

(57) A phoneme recognition system includes a parameter generator for generating a plurality of acoustic parameters including a transient detection parameter, Fig 5A, corresponding to an input voice signal, Fig 5G, a detector for detecting feature points of the acoustic parameters, a generator for generating the difference between two adjacent transient detection parameters, and detectors for detecting stationary and transient parts of the input voice signal according to the generated difference so as to provide more precise phoneme segmentation.

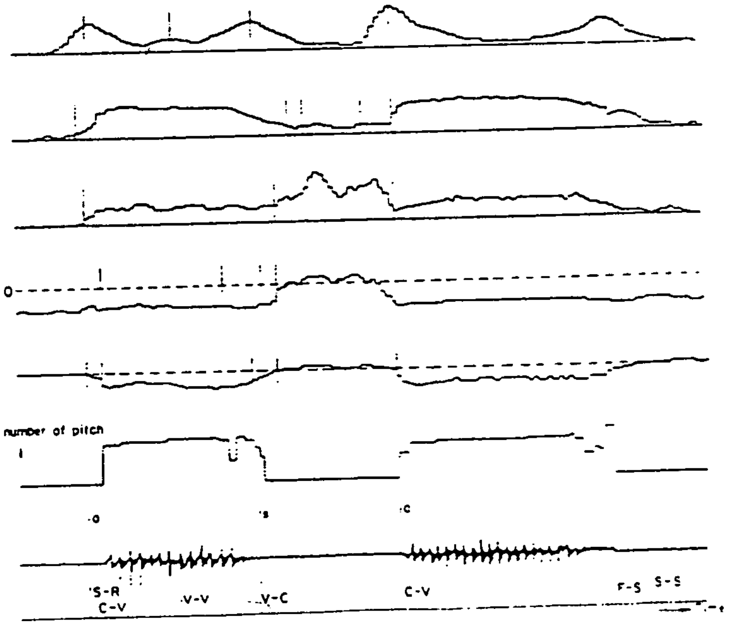


FIG. 5A

FIG. 5B

FIG. 5C

FIG. 5D

FIG. 5E

FIG. 5F

FIG. 5G

GB 2 213 623 A

FIG. 1

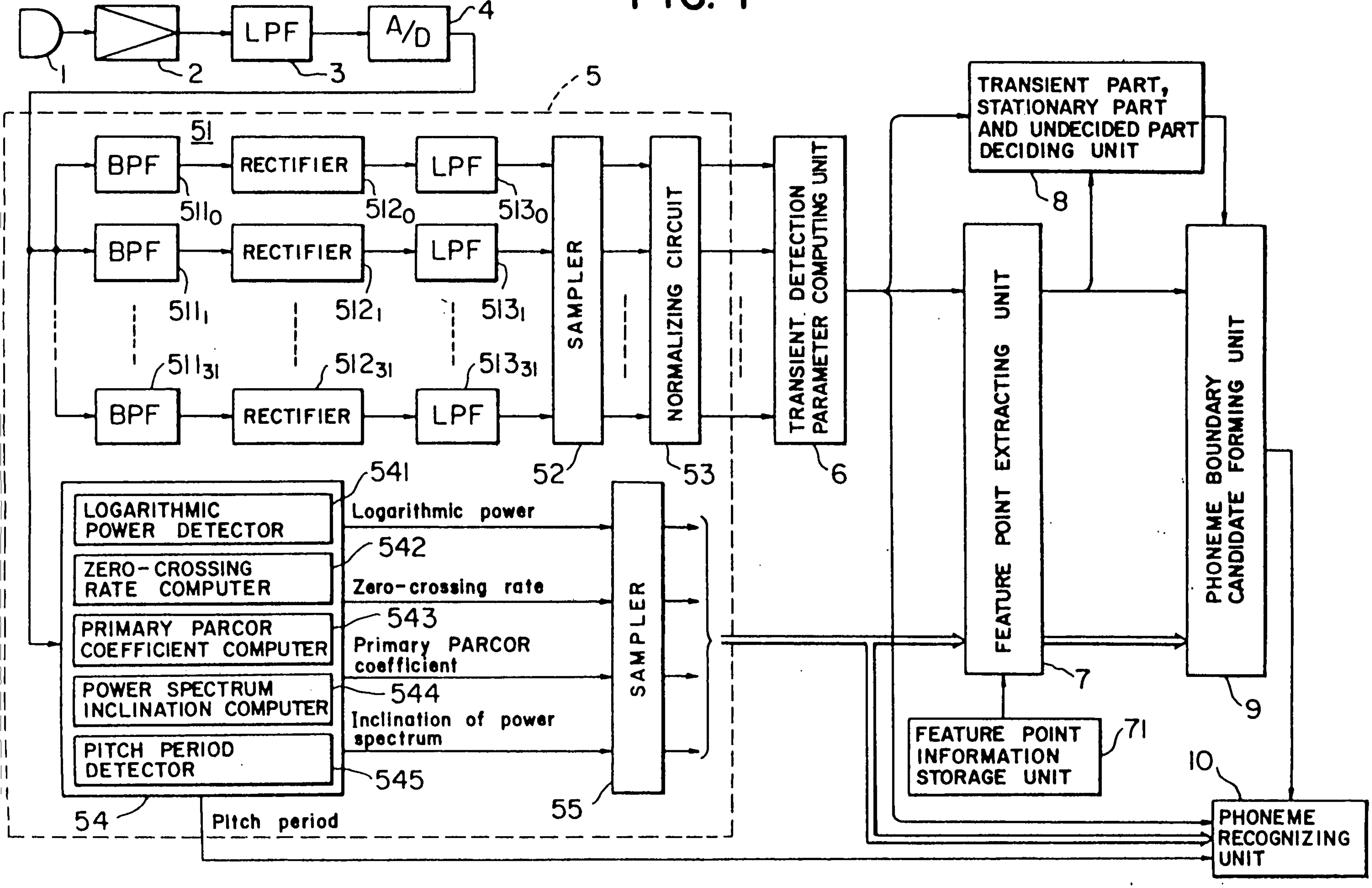

1
2
LPF
3
A/D
4
5
51
BPF
511₀
RECTIFIER
512₀
LPF
513₀
BPF
511₁
RECTIFIER
512₁
LPF
513₁
511₃₁
BPF
512₃₁
RECTIFIER
513₃₁
LPF
SAMPLER
52
NORMALIZING CIRCUIT
53
TRANSIENT DETECTION PARAMETER COMPUTING UNIT
6
TRANSIENT PART, STATIONARY PART AND UNDECIDED PART DECIDING UNIT
8
FEATURE POINT EXTRACTING UNIT
7
PHONEME BOUNDARY CANDIDATE FORMING UNIT
9
LOGARITHMIC POWER DETECTOR
541
Logarithmic power
ZERO-CROSSING RATE COMPUTER
542
Zero-crossing rate
PRIMARY PARCOR COEFFICIENT COMPUTER
543
Primary PARCOR coefficient
POWER SPECTRUM INCLINATION COMPUTER
544
Inclination of power spectrum
PITCH PERIOD DETECTOR
545
54
Pitch period
SAMPLER
55
FEATURE POINT INFORMATION STORAGE UNIT
71
10
PHONEME RECOGNIZING UNIT

FIG.2

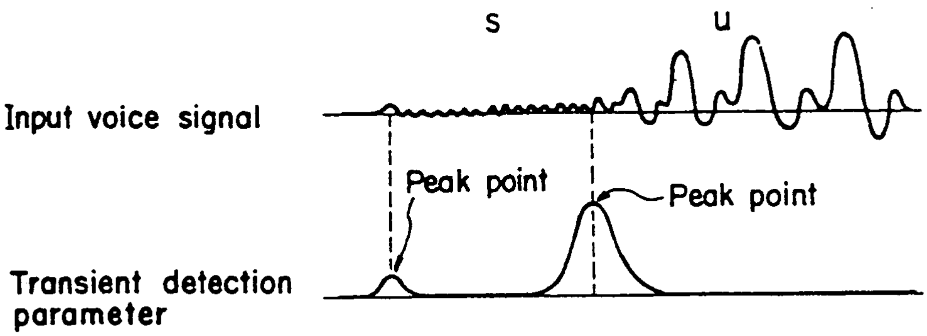


FIG.3

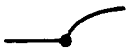
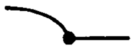
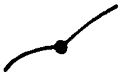
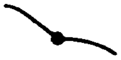
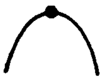
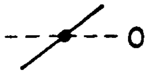
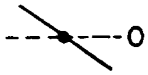
① Rising point	
② Falling point	
③ Increasing turning point	
④ Decreasing turning point	
⑤ Peak point	
⑥ Positive zero-crossing point	
⑦ Negative zero-crossing point	

FIG. 4

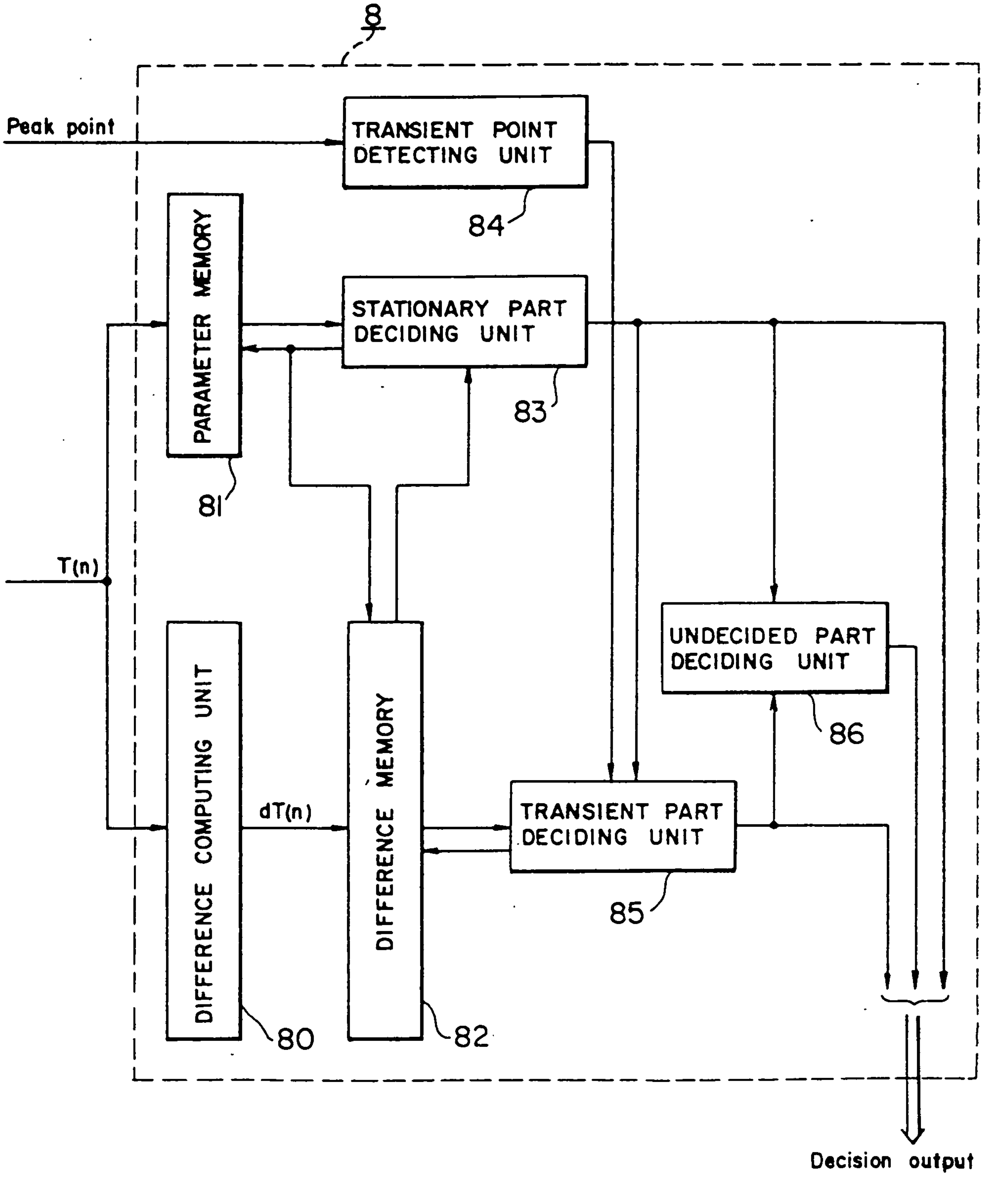

8
Peak point
TRANSIENT POINT DETECTING UNIT
84
PARAMETER MEMORY
STATIONARY PART DECIDING UNIT
83
81
T(n)
UNDECIDED PART DECIDING UNIT
86
DIFFERENCE COMPUTING UNIT
dT(n)
DIFFERENCE MEMORY
TRANSIENT PART DECIDING UNIT
85
80
82
Decision output

FIG. 5A

Transient detection parameter

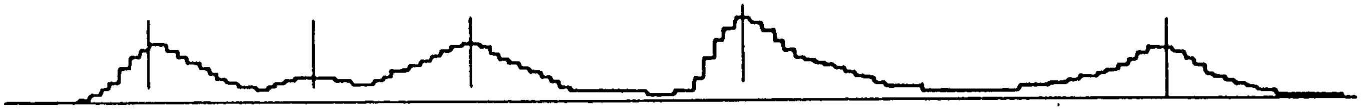


FIG. 5B

Logarithmic power

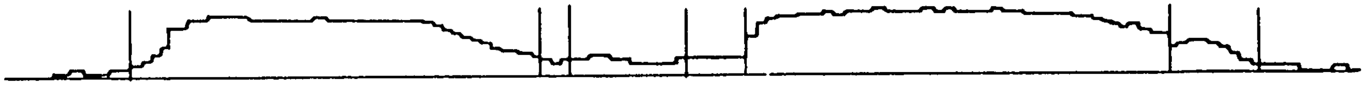


FIG. 5C

Zero-crossing rate

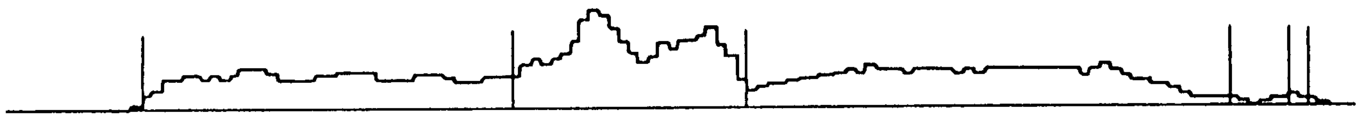


FIG. 5D

Primary PARCOR coefficient

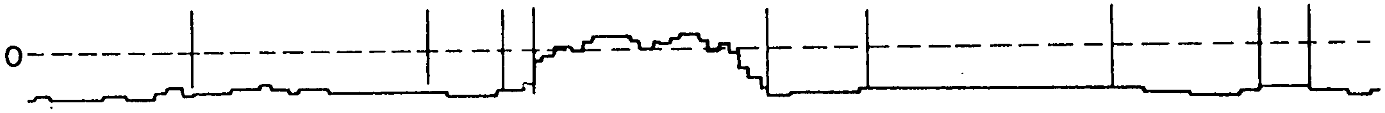


FIG. 5E

Inclination of power spectrum

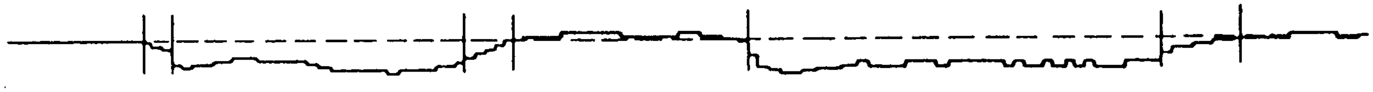


FIG. 5F

Pitch period

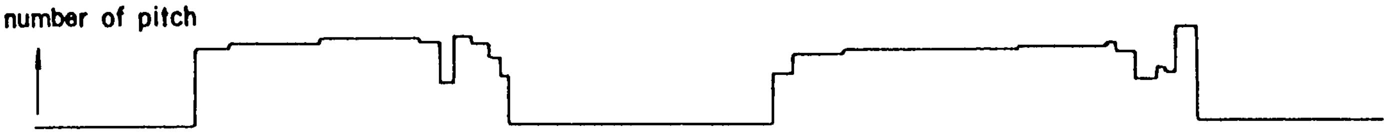


FIG. 5G

Input voice signal

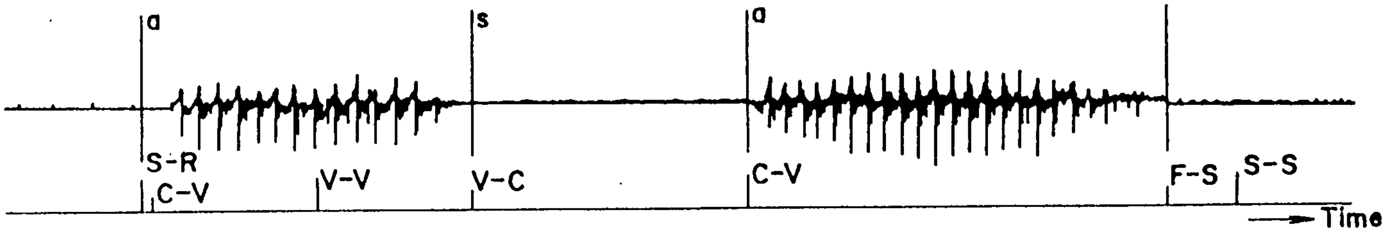


FIG. 6

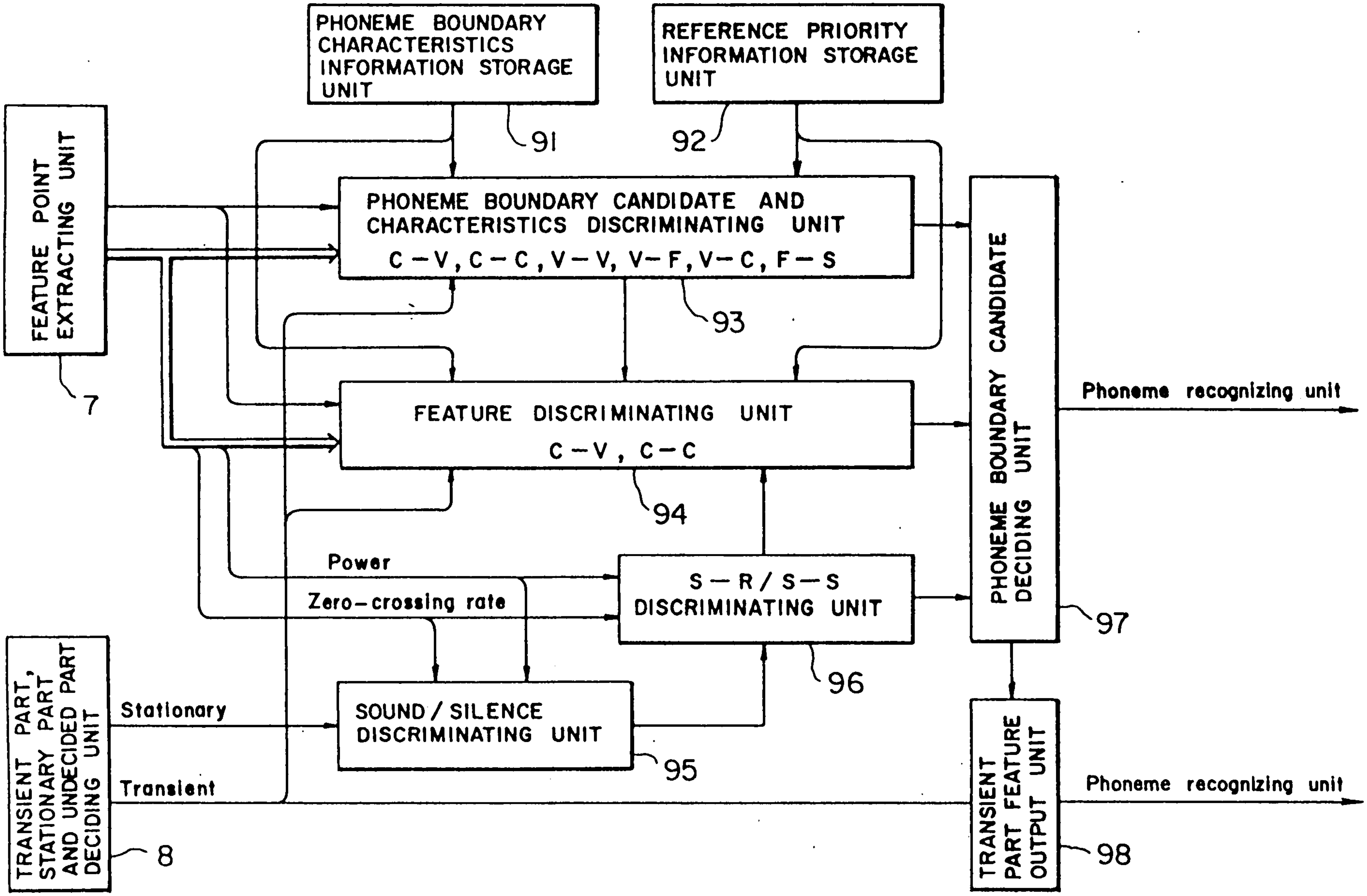

PHONEME BOUNDARY CHARACTERISTICS INFORMATION STORAGE UNIT
91
REFERENCE PRIORITY INFORMATION STORAGE UNIT
92
FEATURE POINT EXTRACTING UNIT
7
PHONEME BOUNDARY CANDIDATE AND CHARACTERISTICS DISCRIMINATING UNIT
C – V, C – C, V – V, V – F, V – C, F – S
93
FEATURE DISCRIMINATING UNIT
C – V, C – C
94
Power
Zero-crossing rate
S – R / S – S DISCRIMINATING UNIT
96
TRANSIENT PART, STATIONARY PART AND UNDECIDED PART DECIDING UNIT
8
Stationary
Transient
SOUND / SILENCE DISCRIMINATING UNIT
95
PHONEME BOUNDARY CANDIDATE DECIDING UNIT
97
Phoneme recognizing unit
TRANSIENT PART FEATURE OUTPUT UNIT
98
Phoneme recognizing unit

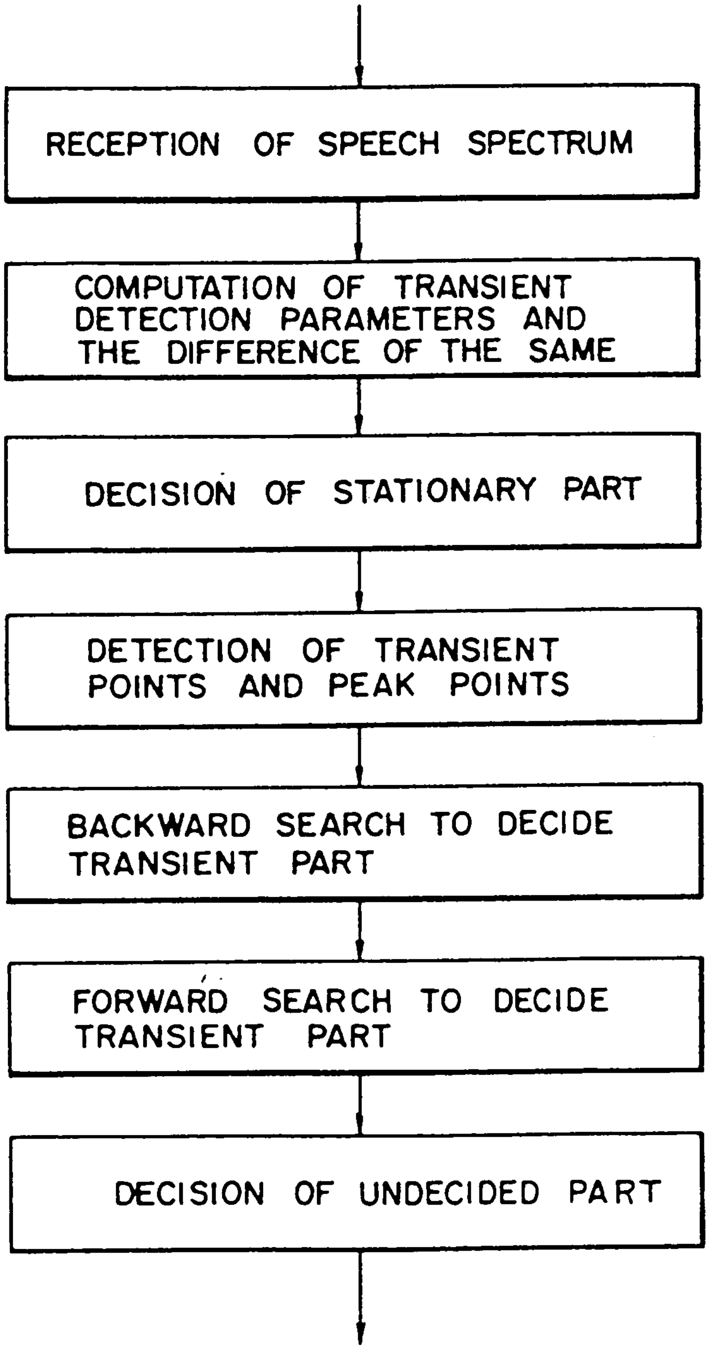
FIG. 7

S — R	Power	①
	Zero-crossing rate	①
C — V	Power	③
	Zero-crossing rate	④
	PARCOR coefficient	④
	Inclination of power spectrum	⑦
¦	¦	¦

FIG. 8

1.	Zero-crossing rate < Power
2.	Inclination of power spectrum < PARCOR coefficient < Zero-crossing rate < Power
3.	Inclination of power spectrum < Zero-crossing rate < Power
¦	¦

FIG. 9


RECEPTION OF SPEECH SPECTRUM
COMPUTATION OF TRANSIENT DETECTION PARAMETERS AND THE DIFFERENCE OF THE SAME
DECISION OF STATIONARY PART
DETECTION OF TRANSIENT POINTS AND PEAK POINTS
BACKWARD SEARCH TO DECIDE TRANSIENT PART
FORWARD SEARCH TO DECIDE TRANSIENT PART
DECISION OF UNDECIDED PART

PHONEME RECOGNITION AND VOICE SIGNAL STATUS DETECTION SYSTEMS

This invention relates to phoneme recognition and voice signal status detection systems, which may form phoneme segment information for segmenting input speech into phoneme segments, particularly for phoneme recognition in speech recognition.

Phoneme recognition is the basis of the recognition of continuous speech and large vocabulary speech. Objective input speech must be segmented into phoneme segments for phoneme recognition. For example, when a syllable "SU" is pronounced, the sound waveform can be segmented into a phoneme segment of the consonant "S" and a phoneme segment of the vowel "U".

A method of obtaining a segment boundary by comparing the power or zero-crossing rate of speech with a threshold has been used as a method of phonemic segmentation. However, it has been difficult to achieve accurate phonemic segmentation simply by comparing the power or zero-crossing rate of speech with a threshold, because the setting of the threshold is difficult. According to this method, a transient detection parameter is compared with a threshold to detect a transient part which is greater than the threshold, and a stationary part which is smaller than the threshold. The principal object of the transient detection parameter is the detection of a point where the speech spectrum varies most sharply, namely, a peak point. Therefore, it is difficult to measure the transient state and the stationary state through the simple application of the transient detection parameter. Since it is difficult to set the threshold, it is accordingly difficult to discriminate accurately between the stationary part and the transient part.

According to an aspect of the invention there is provided a system for detecting voice signal status, the system comprising:

sound analysing means arranged to receive an input voice signal for acoustically analysing said input voice signal and for providing a speech spectrum thereof;

means receiving said speech spectrum from said sound analysing means for deriving transient detection parameters;

means receiving said derived transient detection parameters for generating the difference between two successive transient detection parameters;

first detecting means receiving said difference for detecting a stationary part of said input voice signal; and

second detecting means receiving said difference for detecting a transient part of said input voice signal.

According to another aspect of the invention there is provided a system for recognising phoneme boundaries in a voice signal, the system comprising:

means arranged to receive an input voice signal for acoustically analysing said input voice signal and for generating a plurality of acoustic parameters;

means receiving a plurality of said acoustic parameters for detecting feature points of said acoustic parameters;

means for producing phoneme segment boundary candidates and phoneme boundary characteristics corresponding to each of said phoneme segment boundary candidates, based on said detected feature points; and

means for recognising phoneme boundaries in said input voice signal according to said phoneme segment boundary candidates and said phoneme boundary characteristics corresponding thereto.

A preferred embodiment of the present invention, to be described in greater detail hereinafter, provides a phoneme recognition system capable of detecting the stationary part, transient part, and an undecided part, namely neither the stationary part nor the transient part, of input speech, at higher accuracy from transient detection parameters each equivalent to the sum of variance of frequency channels within a block on the time axis, and the difference between the transient detection parameters.

The preferred embodiment provides a phoneme recognition system capable of forming phoneme segment information by obtaining phoneme segment boundary candidates from the feature point information including the rising points, falling points and peak points of a plurality of phoneme segment parameters obtained through the sound analysis of input speech, and the phoneme boundary features of boundary candidates including a rising from silence phoneme, and transition of consonant-to-vowel and vowel-to-vowel, and which is capable of discriminating the phoneme segment accurately and efficiently on the basis of the peak feature points of the transient detection parameters, by using each transient detection parameter as one of the phoneme segment parameters.

According to a further aspect of the invention there is provided a system comprising a sound analyser for acoustically analysing an input voice signal and for providing a speech spectrum thereof, a first generator for generating a transient detection parameter from the speech spectrum, a second generator for generating the difference between two adjacent transient detection parameters, and a detector for detecting stationary portions and transient portions of the input voice signal according to the generated difference between the adjacent two of the transient detection parameters.

The system may further comprise another generator for generating a plurality of acoustic parameters of the input voice signal such as a logarithm power spectrum and a zero cross rate, another detector for detecting feature points of the acoustic parameters such as rising points, increasing points, peak points and the like, so that the system may provide phoneme segment boundary features of the input voice signal.

The invention will now be described by way of example with reference to the accompanying drawings, throughout which like parts are referred to by like references, and in which:

Figure 1 is a block diagram of a phoneme recognition system according to an embodiment of the invention;

Figure 2 illustrates an example of a waveform of an input voice signal and a transient detection parameter corresponding thereto;

Figure 3 shows examples of feature points;

Figure 4 is a more detailed block diagram of a stationary and transient portion detector shown in Figure 1;

Figure 5 shows examples of waveforms of an input voice signal and acoustic parameters thereof;

Figure 6 is a more detailed block diagram of a phoneme boundary candidate generator shown in Figure 1;

Figure 7 is a table showing the relationship of phoneme boundary characteristics to acoustic parameters and feature points;

Figure 8 shows the relative priorities of each acoustic parameter; and

Figure 9 is a flow chart of the operation for detecting the stationary and transient portions of the input voice signal.

A phoneme recognition system in accordance with an embodiment of the present invention obtains phoneme segment information on the basis of the peak feature points of transient detection parameters. Prior to description of the phoneme recognition system, the transient detection parameters will be explained.

As an example, when the syllable "SU" is pronounced, a voice waveform as shown in Figure 2A is obtained. Thus, the syllable "SU" can be phonemised into a consonant "S" and a vowel "U". As can be seen from the voice waveform, a phoneme boundary exists in a transient part of the voice waveform where the phoneme changes. Accurate phoneme recognition can be achieved by recognising the phoneme in a stationary part of a phoneme segment.

Use of transient detection parameters is an effective means of detecting the transient state and the stationary state.

The transient detection parameter is represented by the variation of a speech spectrum defined by the sum of variance within a block on the time axis of each frequency channel.

That is, first the gain of the speech spectrum Si(n) is normalised by the average Savg(n) in the direction of frequency.

$$\mathrm{Savg}(n) = \sum_{i=1}^{q} \mathrm{Si}(n)/q \qquad \ldots\ldots (1)$$

where i is the channel number, and q is the number of channels. The information relating to each of the q channels is sampled on the time axis. A block of information concerning the q channels at the same time point is designated a frame. In the expression (1), n is the frame number of a frame for recognition.

The gain-normalised voice spectrum $\widehat{\mathrm{Si}}(n)$ is expressed by

$$\widehat{\mathrm{Si}}(n) = \mathrm{Si}(n) - \mathrm{Savg}(n) \qquad \ldots\ldots (2)$$

A transient detection parameter T(n) is represented by the sum of variance on the time axis of each channel within blocks (n-M, n+M) which are the sum (2M+1) of M frames before and after the frame.

$$T(n) = \sum_{i=1}^{q} \sum_{j=-M}^{M} \left|\hat{S}i(n+j) - Ai(n)\right|^{a} \qquad \ldots\ldots (3)$$

$$Ai(n) = \sum_{j=-M}^{M} \hat{S}i(n+j)/(2M+1) \qquad \ldots\ldots (4)$$

where Ai(n) is the average on the time axis within the block of each channel.

In particular, since the variation in the central part of the [n-M, n+M] block is liable to pick up fluctuation in sound and noise, the expression (3) is developed into an expression (5) to eliminate the variation in the central part in calculating the transient detection parameter T(n).

$$T(n) = \frac{\{\sum_{i=1}^{q} \sum_{j=-M}^{-M} \left|\hat{S}i(n+j) - Ai(n)\right|^{a} + \sum_{j=M}^{M} \left|\hat{S}i(n+j) - Ai(n)\right|^{a}\} 2q(M-m+1)}{2q(M-m+1)} \qquad \ldots\ldots (5)$$

The transient detection parameter T(n) may be determined, for example, by substituting a = 1, M = 28, m = 3 and q = 32 into the expression (5). In the case of the input speech "SU", a transient detection parameter as shown in Figure 2B is obtained.

The peak points of the transient detection parameter T(n) are stable features in the transient parts. Determination of phoneme boundary candidates on the basis of the transient detection parameters T(n) enables the avoidance of erroneous selection of phoneme boundary candidates. The system embodying the present invention particularly utilises such characteristics of the transient detection parameters.

Figure 1 shows a phoneme recognition system according to a preferred embodiment of the present invention, which is equipped with a phoneme segment information forming apparatus. A speech signal generated by a microphone 1 is transmitted through an amplifier 2 and a low-pass filter 3 for limiting bandwidth to an analog-to-digital (A/D) converter 4 which samples the speech signal, for example, at a sampling frequency of 12.5 kHz to convert the analog speech signal into a digital speech signal which is then supplied to a sound analysing unit 5.

The sound analysing unit 5 comprises a band pass filter bank 51 and a sound analyser 54. The band pass filter bank 51 may comprise, for example, thirty-two channels of digital band pass filters 511_0, 511_1, 511_2, ... 511_{31}. The digital band pass filters 511_0, 511_1, ... and 511_{31} may, for example, be Butterworth digital filters of fourth degree having equal division bands of a bandwidth between 250 Hz and 5.5 kHz on a logarithmic axis. The output signals of the digital band pass filters 511_0, 511_1 ... and 511_{31} are applied to rectifiers 512_0, 512_1, ... and 512_{31}, respectively. The output signals of the rectifiers 512_0, 512_1, ... 512_{31} are applied to digital low-pass filters 513_0, 513_1, ... 513_{31}, respectively. The digital low-pass filters 513_0, 513_1, ... 513_{31} may, for example, be FIR low-pass filters having a cutoff frequency of 52.8 Hz. The output signals of the digital low-pass filters 513_0, 513_1, ... and 513_{31} are applied to a sampler 52. The sampler 52 samples the output signals of the digital low-pass filters 513_0, 513_1, ... and 513_{31} at a frame period of 5.12 ms. Thus a sample time series forming the speech spectrum Si(n) (i = 1, 2, ... and 32, n = 1, 2, ... and N (frame number) 0) is obtained.

The output signal of the sampler 52, namely, the sample time series Si(n), is applied to a normalisation circuit 53 to obtain a time series Si(n) of a normalised speech spectrum.

The sample time series Si(n) of the speech spectrum provided by the normalisation circuit 53 is applied to a transient detection parameter computing unit 6, which executes computation by using the expression (5) to obtain the transient detection parameters T(n). In the computation using the expression (5), for example, M = 5 and m = 2 (which are smaller than M = 28 and m = 3 used in the foregoing computation) may be used to detect the transient parts and the stationary parts and to reduce the number of computations.

The transient detecting parameter T(n) for an input speech "ASA" for instance, is shown in Figure 5A. Figure 5G is the waveform of the input speech signal.

The sound analyser 54 of this embodiment comprises a logarithmic power detector 541 for detecting the logarithmic power of the input speech signal, a zero-crossing rate computer 542, a computer 543 for computing a primary PARCOR coefficient indicating the degree of correlation between the successive samples, a computer 544 for computing the inclination of the power spectrum, and a pitch period detector 545 for detecting the pitch period of the input speech signal. The detected pitch period is applied to a phoneme recognising unit 10.

In the computation of these parameters, namely, the logarithmic power, the zero-crossing rate, the primary PARCOR coefficient, the inclination of power spectrum and the pitch period, a window having a time width corresponding to M frames before a time point (frame) and M frames after the time point is shifted successively by one sampling point at a time on the time axis to generate the parameters by carrying out computation within each window. These parameters are given to a sampler 55, which samples the parameters at the same sampling pulses as those for the sampler 52. Accordingly, the sampler 55 provides parameters of analysed information in the same time series as that for the speech spectrum Si(n).

Figures 5B, 5C, 5D and 5E respectively show the logarithmic power, the zero-crossing rate, the primary PARCOR coefficient and the inclination of the power spectrum thus obtained. Figure 5F shows the pitch period of the speech.

The parameters thus obtained by the sound analysing unit 5 are fed as parameters for a recognising process to the phoneme recognising unit 10. The transient detection parameters T(n) computed by the transient detection parameter computing unit 6 and the parameters determined by the sound analysing unit 54 excluding the pitch period are fed to a feature point extracting unit 7.

The feature point extracting unit 7 extracts general feature points to obtain phoneme boundary candidates from the parameters for segmentation. In this example, the following seven feature points (1) to (7) as shown in Figure 3 are used.

(1) Rising point.

(2) Falling point.

(3) Increasing turning point.

(4) Decreasing turning point.

(5) Peak point.

(6) Positive zero-crossing point.

(7) Negative zero-crossing point.

The feature point extracting unit 7 extracts the feature points of the parameters with reference to feature point information provided by a feature point information storage unit 71. In Figures 5A to 5E, positions on the time axis indicated by vertical lines are the feature points of the parameters. For example, peak points (5) may be extracted as the feature points of the transient detection parameters T(n), and rising points (1), falling points (2), increasing turning points (3) and decreasing turning points (4) may be extracted as the feature points of the parameters of the logarithmic power and the zero-crossing rate.

The feature point information obtained by the feature point extracting unit 7 is applied to a phoneme boundary candidate forming unit 9, which determines phoneme boundary candidates on the basis of the transient detection parameters T(n) and extracts the features of the phoneme boundary candidates.

The phoneme boundary candidate forming unit 9 makes reference to a decision output provided by a transient part, stationary part and undecided part deciding unit 8. The deciding unit 8 receives the transient detection parameters T(n) from the transient detection parameter computing unit 6, and the peak feature point information on the transient detection parameters T(n) from the feature point extracting unit 7, and then the deciding unit 8 decides which are the undecided parts belonging to neither the transient parts of the input speech nor the stationary parts of the input speech.

The transient part, stationary part and undecided part deciding unit 8 is shown in Figure 4 as comprising a difference computing unit

80, a parameter memory 81, a difference memory 82, a stationary part deciding unit 83, a transient point detecting unit 84, a transient part deciding unit 85 and an undecided part deciding unit 86.

The transient detection parameters T(n) provided by the transient detection parameter computing unit 6 are applied to the difference computing unit 80 to compute the difference dT(n) between the successive transition detection parameters.

$$dT(n) = T(n+1) - T(n) \qquad \ldots\ldots (6)$$

The parameter memory 81 stores the transient detection parameters T(n) provided by the transient detection parameter computing unit 6, and the difference memory 82 stores the difference dT(n).

The deciding operation will now be described.

(i) The stationary part deciding unit 83 sends a search signal to the memories 81 and 82 to read the transient detection parameters T(n) and the difference dT(n) sequentially from the memories 81 and 82, and decides a segment to be a stationary part when the segment meets the conditions

$$T(n) \leqq T_{s1} \qquad \ldots\ldots (7)$$

or

$$T(n) \leqq T_{s2} \ (T_{s1} < T_{s2})$$

and

$$dT(n) \leqq |d_0| \qquad \ldots\ldots (8)$$

where T_{s1}, T_{s2} and d_0 are set thresholds, for example, $T_{s1} = 1.0$, $T_{s2} = 1.5$ and $d_0 = 0.1$.

(ii) The transient point detecting unit 84 detects the peak points of the transient detection parameters T(n) (Figure 5B) from the feature point extracting unit 7, regards the peak points as transient points each representing the centre of the transient part, and then gives position information (frame numbers) about the transient points to the transient part deciding unit 85.

(iii) The transient part deciding unit 85 sends a search signal having the basic point on the transient point to the difference memory 82 to read the difference dT(n). The past difference is searched backwards with respect to time from the transient point as a basic point (hereinafter, this mode of search will be referred to as "backward search") and decides a segment having a difference dT(n) meeting the condition

$$dT(n) \geq d_1 \quad (d_1 \text{ is a threshold}) \qquad \ldots\ldots (9)$$

to be a rear transient part. For example, $d_1 = 0.2$.

(iv) In the backward search, when a segment meeting the expression (9) overlaps a stationary part decided by the stationary part deciding unit 83, a segment immediately before a portion of the segment before the stationary part is regarded as a transient part.

(v) Then, the transient part deciding unit 85 makes a search forwards with respect to time (hereinafter, this mode of search is referred to as "forward search") from the transient point as a basic point and decides a segment having a dT(n) meeting an inequality

$$dT(n) \leq -d_1 \qquad \ldots\ldots (10)$$

to be a forward transient part.

(vi) In the forward search, when the segment meeting the expression (10) overlaps a stationary part, a portion of the segment immediately before the stationary part is regarded as a transient part.

(vii) A transient part having its centre on a transient point is detected from the backward transient part and the forward transient part. The foregoing procedure is executed for all the transient points to discriminate all the transient parts.

(viii) Then, the undecided part deciding unit 86 makes reference to the respective decision outputs of the stationary part deciding unit 83 and the transient part deciding unit 85 and decides segments decided to be neither a stationary part nor a transient part to be undecided parts. in Figure 5A, parts indicated by thick solid lines are transient parts, parts indicated by thin solid lines are stationary parts, and parts indicated by broken lines are undecided parts.

The decision output of the undecided part deciding unit 86 is supplied, together with the respective decision outputs of the stationary part deciding unit 83 and the transient part deciding unit 85, to the phoneme boundary candidate forming unit 9.

Attention is directed particularly to the stationary parts among data included in the decision output of the deciding unit 8 applied to the phoneme recognising unit 10 for phoneme recognition, and the undecided parts are ignored to achieve accurate phoneme recognition, because the undecided parts are factors of variation. A computer may be employed for carrying out the foregoing operation. Figure 9 is a flow chart showing procedures for deciding the stationary part, the transient part and the undecided part.

The phoneme boundary candidate forming unit 9 will be described hereinafter with reference to Figure 6.

The phoneme boundary candidate forming unit 9 determines phoneme boundary candidates. The following eight phoneme boundary characteristics are used.

(1) Rise from silence (S-R).

(2) Consonant-to-vowel transition (C-V).

(3) Consonant-to-consonant transition (C-C).

(4) Vowel-to-vowel transition (V-V).

(5) Fall-to-vowel transition (V-F).

(6) Vowel-to-consonant transition (V-C).

(7) Fall-to-silence transition (F-S).

(8) Sound-to-silence transition (S-S).

A phoneme boundary characteristics information storage unit 91 stores data representing these eight phoneme boundary characteristics. A phoneme boundary candidate and characteristics discriminating unit 93 discriminates phoneme boundary characteristics of phoneme boundary candidates with reference to information fetched from the phoneme boundary characteristics information storage unit 91. In Figure 7, the phoneme boundary characteristics data are represented by the symbols S-R, C-C, C-V and the like. Also shown in Figure 7 are sound parameters constituting phoneme boundaries, and the numbers (1), (2), (3), ... of the feature points extracted by the feature point extracting unit 7 shown in Figure 3. Each of the phoneme boundary characteristics may correspond to the plurality of sound parameters and feature points.

A reference priority information storage unit 92 stores reference priority information of the sound parameters as shown in Figure 8, in which the priority of the right-hand parameter is higher than that of the left-hand parameter.

A phoneme boundary candidate and characteristics discriminating unit 93 collects the feature points of the parameters to decide a phoneme boundary candidate and determines the phoneme boundary characteristics of the phoneme boundary candidate using the feature points obtained by the feature point extracting unit 7 dislocated or undetected with the parameters.

In this operation, the discriminating unit 93 makes reference to the transient part decision output provided by the deciding unit 8. The discriminating unit 93 regards the transient point in the transient part, namely, the peak feature point of the transient detection parameter, as the first phoneme boundary candidate, and examines the feature point of other sound parameter in the vicinity of the transient point to determine a phoneme boundary candidate. In this operation, the discriminating unit 93 decides the reference priority of each parameter with reference to the reference priority information provided by the storage unit 92, and discriminates a phoneme boundary feature corresponding to the feature point of the sound parameter regarded as the phoneme boundary candidate with reference to the phoneme boundary characteristics information provided by the memory unit 91.

Thus, the discriminating unit 93 discriminates the phoneme boundary characteristics of C-V, C-C, V-V, V-F, V-C and F-S.

Another feature discriminating unit 94 makes reference to the transient part decision output provided by the deciding unit 8 to search for further feature points before the transient point other than the phoneme boundary candidate discriminated by the discriminating unit 93. If any feature point is found, the feature discriminating unit 94 discriminates the phoneme boundary characteristics of C-V and C-C by using the detected feature point. The discriminating unit 94 deals with the following cases. For example, a bilabial voiced plosive "BA" has little stationary part between transient parts, the two transient parts being close to each other, and hence only one feature point can be detected. Therefore, a peak feature point which must originally be before the transient point is detected from the feature point of another parameter. A feature point after the transient point is not searched for, because, in the Japanese language, a vowel is preceded by a consonant, and the peak of the vowel is higher than that of the consonant.

Naturally, different languages differ in the expected position of a feature point to be searched for, and hence a method suitable for the specific language is applied to searching for the feature point.

A sound/silence discriminating unit 95 receives the stationary part decision output of the deciding unit 8, and discriminates between the stationary part of sound and the stationary part of silence from the feature point information about the logarithmic power and the zero-crossing rate.

An S-R/S-S discriminating unit 96 receives the sound/silence discrimination output of the sound/silence discriminating unit 95 and the feature point information about the logarithmic power and the zero-crossing rate, and then discriminates between the phoneme boundary feature S-R of rise from silence and the phoneme boundary feature S-S of transition from sound to silence.

The results of discrimination of the discriminating units 93, 94 and 96 are given to a phoneme boundary candidate deciding unit 97. Then, the phoneme boundary candidate deciding unit 97 applies collectively the position (frame) of the phoneme boundary candidate and the phoneme boundary characteristics obtained by the discriminating units 93, 94 and 96 to the phoneme recognising unit 10. The phoneme boundary candidate and the phoneme boundary features of the specific example are shown under the speech waveform shown in Figure 5G.

In this example, a transient part feature output unit 98 receives the phoneme boundary characteristics from the phoneme boundary candidate deciding unit 97, and the transient part decision output from the deciding unit 8. Then, the transient part characteristics output unit 98 gives a phoneme boundary characteristic of the transient part including the phoneme boundary to the phoneme recognising unit 10.

The phoneme recognising unit 10 carries out phoneme recognition by using the parameters provided by the sound analysing unit 5 and making reference to the phoneme segment information provided by the phoneme boundary candidate forming unit 9. Then, the phoneme recognising unit 10 determines a phoneme symbol and gives the phoneme symbol, for example, to a continuous speech and large vocabulary speech recognising unit (not shown).

The hardware part of this embodiment, namely, the feature point extracting unit 7, the transient part, stationary part and undecided part deciding unit 8, the phoneme boundary candidate forming unit 9, and the operating elements of the sound analysing unit 5, may be substituted by computer software.

Thus, the phoneme recognition system extracts feature points expected to be phoneme boundaries from a plurality of parameters obtained through sound analysis, and decides a phoneme segment candidate from the data of the feature points of the plurality of parameters. Accordingly, more accurate phoneme segment information can readily be obtained. Furthermore, since the phoneme segment information includes the characteristics of the phoneme segment candidate, phoneme recognition can easily be achieved.

Further, since the phoneme boundary candidate is decided on the basis of the peak point of the transient detection parameter, which is a stable feature point in a transient part of the input speech, the selection of an erroneous phoneme boundary candidate is obviated.

Since the difference between the transient detection parameters is calculated, the stationary part is decided on the basis of the transient detection parameters and the difference, and the transient part is decided on the basis of the difference, instead of deciding the stationary part and the transient part through the simple comparison of the transient detection parameters with a threshold; accurate decision of the stationary part and the transient part is thus achieved.

Furthermore, since the system identifies a segment which is neither a stationary part nor a transient part to be an undecided part, phoneme segment decision and phoneme recognition can be achieved by using segments excluding the undecided parts, which are the factors of variation, and the undecided part decision output.

CLAIMS

1. A system for detecting voice signal status, the system comprising:

sound analysing means arranged to receive an input voice signal for acoustically analysing said input voice signal and for providing a speech spectrum thereof;

means receiving said speech spectrum from said sound analysing means for deriving transient detection parameters;

means receiving said derived transient detection parameters for generating the difference between two successive transient detection parameters;

first detecting means receiving said difference for detecting a stationary part of said input voice signal; and

second detecting means receiving said difference for detecting a transient part of said input voice signal.

2. A system according to claim 1, comprising third detecting means receiving output signals from said first detecting means and said second detecting means for detecting an undecided part of said input voice signal, said undecided part being neither said stationary part nor said transient part.

3. A system according to claim 1 or claim 2, comprising peak detecting means for detecting a peak of said transient detection parameters.

4. A system for detecting voice signal status, the system being substantially as herein described with reference to the accompanying drawings.

5. A system for recognising phoneme boundaries in a voice signal, the system comprising:

means arranged to receive an input voice signal for acoustically analysing said input voice signal and for generating a plurality of acoustic parameters;

means receiving a plurality of said acoustic parameters for detecting feature points of said acoustic parameters;

means for producing phoneme segment boundary candidates and phoneme boundary characteristics corresponding to each of said phoneme segment boundary candidates, based on said detected feature points; and

means for recognising phoneme boundaries in said input voice signal according to said phoneme segment boundary candidates and said phoneme boundary characteristics corresponding thereto.

6. A system according to claim 5, wherein said plural acoustic parameters include transient detection parameters and said feature point detecting means is operable to detect a peak point of a said transient detection parameter as one of said feature points, said producing means being operable to produce said phoneme segment boundary candidates and phoneme boundary characteristics corresponding thereto only within a predetermined period from said detected peak point.

7. A system for recognising phoneme boundaries in a voice signal, the system being substantially as herein described with reference to the accompanying drawings.